Volume 108, Number 6, November-December 2003
Journal of Research of the National Institute of Standards and Technology

[J. Res. Natl. Inst. Stand. Technol. **108**, 439-446 (2003)]

Statistical Interpretation of Key Comparison Reference Value and Degrees of Equivalence

Volume 108 Number 6 November-December 2003

R. N. Kacker, R. U. Datla, and A. C. Parr

National Institute of Standards and Technology, Gaithersburg, MD 20899-0001, USA

raghu.kacker@nist.gov
raju.datla@nist.gov
albert.parr@nist.gov

Key comparisons carried out by the Consultative Committees (CCs) of the International Committee of Weights and Measures (CIPM) or the Bureau International des Poids et Mesures (BIPM) are referred to as CIPM key comparisons. The outputs of a statistical analysis of the data from a CIPM key comparison are the key comparison reference value, the degrees of equivalence, and their associated uncertainties. The BIPM publications do not discuss statistical interpretation of these outputs. We discuss their interpretation under the following three statistical models: nonexistent laboratory-effects model, random laboratory-effects model, and systematic laboratory-effects model.

Keywords: interlaboratory evaluation; measurement uncertainty; variance components.

Accepted: February 17, 2004

Available online: http://www.nist.gov/jres

1. Introduction

Key comparisons are interlaboratory comparisons that serve as technical bases for Mutual Recognition Arrangements (MRA) between national metrology institutes (NMIs) [1]. Key comparisons carried out by the Consultative Committees (CCs) of the International Committee of Weights and Measures (CIPM) or the Bureau International des Poids et Mesures (BIPM) are referred to as CIPM key comparisons. Key comparisons carried out by regional metrology organizations (RMO) are referred to as RMO key comparisons. The guidelines for carrying out CIPM key comparisons are given in reference [2].

The objectives of a CIPM key comparison are described in reference [1]. We consider two interpretations of these objectives. A common interpretation is summarized by Nielsen [3] as follows: "The purpose of measurement intercomparisons between NMIs is to test, whether measurements performed in the participating countries are consistent taking into account the uncertainties assigned to the measurements. If an inconsistency is detected, the participating countries should take the corrective actions needed to obtain consistency. Otherwise, measurement results exchanged across borders cannot be considered equivalent without adding a 'between countries uncertainty,' which would be in disharmony with the concept of the SI system of units."

This paper is based on a second interpretation of the objectives of a CIPM key comparison: Generally, the participants of a CIPM key comparison are NMIs that are members of the appropriate Consultative Committee; at least some of these NMIs provide realizations of the SI values to establish the traceability of measurements made in their countries. The purpose

of a CIPM key comparison is to establish the key comparison reference value[1], the degrees of equivalence[2], and their associated uncertainties on the basis of the data provided by the participants.

This paper is limited to a simple CIPM key comparison where the common measurand is a physical quantity of stable value during the comparison. Many CIPM key comparisons are not simple because it is often impractical or impossible to realize exactly the same measurand for or by all participants. We use the symbol Y for the stable value of the measurand. The data provided by the participants of a simple CIPM key comparison are paired results and standard uncertainties $[x_1, u(x_1)], \ldots, [x_n, u(x_n)]$, where the results $x_1, \ldots, x_n$ are measurements of Y. The outputs of a statistical analysis of these data are the key comparison reference value x_R, the degree of equivalence $d_i = x_i - x_R$ of the result x_i, the degree of equivalence $d_{i,j} = d_i - d_j = x_i - x_j$ of the results x_i and x_j, and their associated standard uncertainties $u(x_R)$, $u(d_i)$, and $u(d_{i,j})$, respectively, for $i, j = 1, 2, \ldots, n$ and $i \neq j$ [1]. The key comparison reference value x_R is an estimate for Y. An estimate for Y is a combined result of measurement determined from the data $[x_1, u(x_1)], \ldots, [x_n, u(x_n)]$.

An understanding of the difference between sampling probability distributions, used in classical (frequentist) statistics, and state-of-knowledge probability distributions, used in Bayesian statistics, is necessary for proper analysis and interpretation of the data from a key comparison. Briefly, they are defined as follows. In classical statistics, the value of the measurand is assumed to be an unknown constant, often called the true value, and each result of measurement is regarded as a realization of a random variable with a sampling distribution. A *sampling distribution* is a probability distribution that describes the relative frequencies of occurrence for all possible results of measurement when the conditions of measurement are hypothesized to be fixed at the intended levels [4]. The metrologist relates the expected values of the sampling distributions for the results of measurement to the value of the measurand. A classical (frequentist) statistical interpretation is a statement that relates the realized measurements to what one might expect if the key comparison could be repeated infinitely many times and throughout these repetitions the hypothesized sampling distributions continued to apply.

In Bayesian statistics, the measurement data are given constants and the value of the measurand is a random variable. A probability distribution for the value of the measurand is a *state-of-knowledge distribution* that describes the degrees of belief for all possible values that could be attributed to the measurand [4]. The belief is based on all available information including current results of measurement and scientific judgment based on prior and other data. Similar state-of-knowledge distributions apply to the other parameters involved in assessing the value of the measurand. A Bayesian interpretation is a statement that represents the state-of-knowledge about the value of the measurand based on state-of-knowledge distributions before measurements are made and a likelihood function conditional on the current measurements [4]. The ISO *Guide* [5] is consistent with a Bayesian interpretation of measurements but not with a classical (frequentist) interpretation [4].

We refer to the results $x_1, \ldots, x_n$ as laboratory results. The laboratory results $x_1, \ldots, x_n$ are regarded as realizations of random variables $x_1, \ldots, x_n$ with sampling distributions[3]. We use the symbols $X_1, \ldots, X_n$ for the expected values $E(x_1), \ldots, E(x_n)$ of the sampling distributions of $x_1, \ldots, x_n$, respectively. We refer to the expected values $X_1, \ldots, X_n$ as the laboratory expected values. We use the symbols $\sigma_1, \ldots, \sigma_n$ for the standard deviations $S(x_1), \ldots, S(x_n)$ of the sampling distributions of $x_1, \ldots, x_n$, respectively. Here $S(x_i)$ is the square root of the variance $V(x_i) = E[x_i - E(x_i)]^2$ of the sampling distribution of x_i for $i = 1, 2, \ldots, n$. The uncertainties $u(x_1), \ldots, u(x_n)$ are statistical estimates of $\sigma_1, \ldots, \sigma_n$, respectively.

References [1] and [2] do not discuss statistical interpretations of the pairs $[x_R, u(x_R)]$, $[d_i, u(d_i)]$, and $[d_{i,j}, u(d_{i,j})]$. A statistical analysis of the data from a key comparison and interpretation of its outputs requires assumptions and models about the relationship between the data $[x_1, u(x_1)], \ldots, [x_n, u(x_n)]$ and the value Y of the measurand. In Sec. 2, we discuss two assumptions, labeled as *Assumption I* and *Assumption II*, about the relationship between the laboratory expected values $X_1, \ldots, X_n$ and Y. Then we discuss two classical statistics models, a nonexistent laboratory-effects model

[1] "Key comparison reference value: the reference value accompanied by its uncertainty resulting from a CIPM key comparison [1]."

[2] "Degree of equivalence of a measurement standard: the degree to which the value of a measurement standard is consistent with the key comparison reference value. This is expressed quantitatively by the deviation from the key comparison reference value and the uncertainty of this deviation. The degree of equivalence between two measurement standards is expressed as the difference between their respective deviations from the key comparison reference value and the uncertainty of this difference [1]."

[3] We use the symbols $x_1, \ldots, x_n$ for both the random variables and their realized values.

and a random laboratory-effects model, based on *Assumption I*. Next, we propose a systematic laboratory-effects model based on *Assumption II*. We describe the key comparison reference value, the degrees of equivalence, and their associated uncertainties determined by each of the three statistical models. In Sec. 3 and 4, we discuss statistical interpretations of the pairs $[x_R, u(x_R)]$, $[d_i, u(d_i)]$, and $[d_{i,j}, u(d_{i,j})]$ under the three statistical models. Our conclusion is given in Sec. 5.

2. Statistical Assumptions and Models for the Relationship Between the Data and the Value of the Measurand

In this section, we discuss statistical assumptions and models for analyzing the data from a simple CIPM key comparison to determine the key comparison reference value, the degrees of equivalence, and their associated uncertainties.

2.1 Assumptions About the Relationship Between the Laboratory Expected Values and the Value of the Measurand

One may either assume that the laboratory expected values $X_1, \ldots, X_n$ are all equal or allow for the possibility that $X_1, \ldots, X_n$ may not be equal.

Assumption I: The expected values $X_1, \ldots, X_n$ are all equal. The *Assumption I* defined so far does not specify the relationship between the results $x_1, \ldots, x_n$ and Y. Therefore, in concert with *Assumption I*, it is generally assumed that the common expected value is equal to Y, i.e., $X_1 = \ldots = X_n = Y$. Under *Assumption I*, the results $x_1, \ldots, x_n$ are subject to intralaboratory variations only.

Assumption II: The expected values $X_1, \ldots, X_n$ may not be equal, i.e., $X_i \neq X_j$ for some $i, j = 1, 2, \ldots, n$ and $i \neq j$. Therefore, not all of $X_1, \ldots, X_n$ may equal the value Y of the measurand. The *Assumption II* defined so far does not specify the relationship between the results $x_1, \ldots, x_n$ and Y. Therefore, in concert with *Assumption II*, it is generally assumed that Y is either somewhere in the range of results $x_1, \ldots, x_n$ or in the vicinity of this range[4] [6]. Under *Assumption II*, the results $x_1, \ldots, x_n$ are subject to both the intralaboratory variations represented by the uncertainties $u(x_1), \ldots, u(x_n)$ and the interlaboratory variation arising from the dispersion of $X_1, \ldots, X_n$ about Y. The differences $(X_1 - Y), \ldots, (X_n - Y)$ are laboratory-effects (biases) due to unrecognized sources of error, denoted by $b_1, \ldots, b_n$, in the results $x_1, \ldots, x_n$. The biases are common to all measurements in a particular laboratory but may be different for different laboratories.

2.2 Assumption About the Uncertainties Submitted by the Participants

The standard uncertainties $u(x_1), \ldots, u(x_n)$ submitted by the participants of a key comparison are estimates obtained by combining various estimated components of uncertainty in determining the value Y of the measurand. A combined standard uncertainty $u(x_i)$ may be unreliable for various reasons. For example, a classical (frequentist) Type A component of $u(x_i)$ calculated from a small number of independent measurements is unreliable[5] [5]. A Type A component of $u(x_i)$ based on unjustified statistical assumptions may be unreliable. A Type B component of $u(x_i)$ based on unreasonable state-of-knowledge distributions may be unreliable. A combined uncertainty $u(x_i)$ determined from an incomplete measurement equation may be an underestimate. The unreliability of estimated uncertainties $u(x_1), \ldots, u(x_n)$ is a component of uncertainty in deter-

[4] If the expected value X_1 were equal to the value Y of the measurand, then according to the ISO *Guide*, the interval $[x_1 \pm 2u(x_1)]$ would represent an approximate range of the plausible values of Y. Likewise, if X_2 were equal to Y then the interval $[x_2 \pm 2u(x_2)]$ would represent an approximate range of the plausible values of Y, and so on for $X_3, X_4, \ldots, X_n$. It follows from *Assumption II* that any one or more of the expected values $X_1, \ldots, X_n$ may be close to or equal to Y; therefore, the total interval consisting of the union of intervals $[x_i \pm 2u(x_i)]$, for $i = 1, 2, \ldots, n$, represents an approximate range of the plausible values of Y. However, most metrologists assign greater belief-probability to the middle than to the ends of the total interval.

[5] The unreliability of a classical (frequentist) estimate of uncertainty arising from a small number of measurements is quantified by degrees of freedom [5].

mining the key comparison reference value x_R, the degrees of equivalence d_i and $d_{i,j}$, and their associated standard uncertainties. In this paper, we do not discuss the additional uncertainty that arises from the unreliability of $u(x_1), \ldots, u(x_n)$.

Classical (frequentist) statistical analyses and interpretations discussed in this paper are based on the assumption that the estimated uncertainties $u(x_1), \ldots, u(x_n)$ are equal to the true standard deviations $\sigma_1, \ldots, \sigma_n$ of the sampling distributions of $x_1, \ldots, x_n$, respectively. Most metrologists make this assumption. For example, the expression $u(x_W) = 1/\sqrt{[\Sigma_i w_i]}$ for the standard deviation of the weighted mean $x_W = \Sigma_i w_i x_i / \Sigma_i w_i$, where $w_i = 1/u^2(x_i)$ for $i = 1, 2, \ldots, n$, requires this assumption.

Statistical analyses based on the ISO *Guide* regard a laboratory expected value X_i as a variable with a state-of-knowledge distribution having expected value x_i and standard deviation $u(x_i)$. Such analyses require the assumption that the estimated uncertainties $u(x_1), \ldots, u(x_n)$ are sufficiently reliable.

2.3 Classical (Frequentist) Statistics Models Based on *Assumption I*

The weighted mean $x_W = \Sigma_i w_i x_i / \Sigma_i w_i$ and the expression $u(x_W) = 1/\sqrt{[\Sigma_i w_i]}$, where $w_i = 1/u^2(x_i)$ for $i = 1, 2, \ldots, n$, are often used as the key comparison reference value x_R and its associated standard uncertainty $u(x_R)$, respectively. The use of x_W as x_R and $u(x_W)$ as $u(x_R)$ is based on the following classical (frequentist) statistics model.

2.3.1 Nonexistent Laboratory-Effects Model

The results are regarded as realizations of the random variables $x_1, \ldots, x_n$, where

$$x_i = Y + e_i \,, \qquad (1)$$

and $e_i = (x_i - Y)$ is the error in x_i for $i = 1, 2, \ldots, n$. In this model, the parameter Y is identified with the value of the measurand and the errors $e_1, \ldots, e_n$ are mutually independently distributed random variables with sampling distributions. The sampling distributions of $e_1, \ldots, e_n$ are generally assumed to be normal (Gaussian). The expected values of $e_1, \ldots, e_n$ are assumed to be zero and the variances of $e_1, \ldots, e_n$ are assumed to be $u^2(x_1), \ldots, u^2(x_n)$, respectively. Under model (1) [represented by Eq. (1)], the expected value $E(x_i)$ is equal to Y and the variance $V(x_i)$ is equal to $u^2(x_i)$, for $i = 1, 2, \ldots, n$. Since the expected values of all results are equal to Y, the model (1) is based on *Assumption I*. In model (1), the results $x_1, \ldots, x_n$ are free of laboratory-effects (biases). Therefore, we refer to it as a nonexistent laboratory-effects model. The best least-squares estimate for the parameter Y of the nonexistent laboratory-effects model (1) is the weighted mean $x_W = \Sigma_i w_i x_i / \Sigma_i w_i$, where $w_i = 1/u^2(x_i)$ for $i = 1, 2, \ldots, n$. The term best least-squares estimate[6] means that the estimate x_W has the smallest variance among all estimates of Y that are both linear functions of the results $x_1, \ldots, x_n$ and have the expected value Y. The standard deviation of the sampling distribution of x_W is $u(x_W) = 1/\sqrt{[\Sigma_i w_i]}$. Thus the key comparison reference value x_R based on model (1) is x_W and $u(x_R)$ is $u(x_W)$. The corresponding degrees of equivalence are $d_i = x_i - x_W$ and $d_{i,j} = x_i - x_j$, for $i, j = 1, 2, \ldots, n$ and $i \neq j$. The uncertainties $u(d_i)$ and $u(d_{i,j})$ are determined from the sampling distributions of $x_1, \ldots, x_n$ and x_R under model (1).

Note 1: When not all uncertainties $u(x_1), \ldots, u(x_n)$ are sufficiently reliable estimates of the true standard deviations $\sigma_1, \ldots, \sigma_n$, the true standard deviation of the sampling distribution of the weighted mean x_W may be larger than the true standard deviation of the sampling distribution of the arithmetic mean x_A. Thus in this case the weighted mean x_W may be an inferior key comparison reference value to the arithmetic mean x_A.

2.3.2 Random Laboratory-Effects Model

The classical statistics model based on *Assumption I* for the situation where the dispersion of results $x_1, \ldots, x_n$ may be more than what can reasonably be attributed to the intralaboratory variances $u^2(x_1), \ldots, u^2(x_n)$ is as follows. The results are regarded as realizations of the random variables $x_1, \ldots, x_n$, where

$$x_i = Y + b_i + e_i \,, \qquad (2)$$

$b_i = (X_i - Y)$ is the laboratory effect (bias) in x_i and $e_i = (x_i - X_i)$ is the intralaboratory error in x_i for $i = 1, 2, \ldots, n$. The classical statistics assumptions to relate the results $x_1, \ldots, x_n$ to Y are as follows: the laboratory biases $b_1, \ldots, b_n$ are regarded as random variables having the same normal sampling distribution with expected value zero and variance $\sigma_b^2 \geq 0$, called *interlaboratory variance*; and $b_1, \ldots, b_n$ are assumed to be mutually independent and independent of the errors $e_1, \ldots, e_n$. The model (2) [represented by Eq. (2)] with these assumptions is referred to as a random laboratory-

[6] Least-squares estimation does not require that the errors $e_1, \ldots, e_n$ and hence $x_1, \ldots, x_n$ have normal distributions.

effects model [7]. Here the term random means that the biases $b_1, \ldots, b_n$ are regarded as random variables with the same sampling distribution that is assumed to be normal with expected value zero and variance σ_b^2. Under the random laboratory-effects model (2), the expected value $E(x_i)$ is equal to Y and the variance $V(x_i)$ is equal to $\sigma_b^2 + u^2(x_i)$ for $i = 1, 2, \ldots, n$. The non-existent laboratory-effects model (1) is a special case of the random laboratory-effects model (2) where $\sigma_b^2 = 0$, which means that the biases $b_1, \ldots, b_n$ are all zero, i.e., $X_1 = \ldots = X_n = Y$.

A popular estimate for the parameter Y of model (2) is the weighted mean[7] $x_W = \Sigma_i w_i x_i / \Sigma_i w_i$, where $w_i = 1/[s_b^2 + u^2(x_i)]$ and s_b^2 is an estimate for σ_b^2. Reference [8] discusses various methods for determining s_b^2. The estimate s_b^2 inflates each of the intralaboratory variances $u^2(x_1), \ldots, u^2(x_n)$ just enough to account for the dispersion of results $x_1, \ldots, x_n$ that is not accounted for by model (1). Under the assumption that the estimated variances $s_b^2 + u^2(x_1), \ldots, s_b^2 + u^2(x_n)$ are regarded as the true variances of the sampling distributions of $x_1, \ldots, x_n$, the best estimate of the parameter Y of model (2) is the weighted mean x_W and the standard uncertainty associated with x_W is $u(x_W) = 1/\sqrt{[\Sigma_i w_i]}$, where $w_i = 1/[s_b^2 + u^2(x_i)]$ for $i = 1, 2, \ldots, n$ [9], [8]. Thus the key comparison reference value x_R based on model (2) is the weighted mean $x_W = \Sigma_i w_i x_i / \Sigma_i w_i$ and uncertainty $u(x_R)$ is $u(x_W) = 1/\sqrt{[\Sigma_i w_i]}$, where $w_i = 1/[s_b^2 + u^2(x_i)]$ for $i = 1, 2, \ldots, n$. The corresponding degrees of equivalence are $d_i = x_i - x_R = x_i - x_W$ and $d_{i,j} = x_i - x_j$, for $i, j = 1, 2, \ldots, n$ and $i \neq j$. The uncertainties associated with the degrees of equivalence are determined from the sampling distributions of $x_1, \ldots, x_n$ and x_R under model (2).

The advantage of model (2) relative to model (1) is that it allows for the possibility that the dispersion of results $x_1, \ldots, x_n$ may be more than what can reasonably be attributed to the intralaboratory variances $u^2(x_1), \ldots, u^2(x_n)$. When the dispersion of $x_1, \ldots, x_n$ is not more than what can reasonably be attributed to $u^2(x_1), \ldots, u^2(x_n)$, the estimate s_b^2 is zero. In that case, model (2) yields the same x_R and $u(x_R)$ as model (1). Therefore, there is no disadvantage to using model (2) in place of model (1).

The random laboratory-effects model (2) of classical statistics is conceptually faulty for the analysis of a CIPM key comparison for the following reasons. First, the participants of a CIPM key comparison are specific NMI laboratories rather than randomly chosen from a large population of laboratories. Therefore, the biases $b_1, \ldots, b_n$ may not be regarded as random variables with the same sampling distribution. Second, the assumption that the sampling distribution of the biases $b_1, \ldots, b_n$ is a normal distribution with expected value zero may not be justified. The next section introduces a new model that does not assume that the biases $b_1, \ldots, b_n$ are random variables with a normal sampling distribution.

2.4 A Model Based on *Assumption II* and the ISO *Guide*

A statistical analysis of the data from a simple CIPM key comparison based on *Assumption II* requires one to account for the uncertainty that arises from the unknown bias in a combined result of measurement that is used as an estimate for Y. Before publication of the ISO *Guide*, there was no generally accepted approach to account for the uncertainty that arises from an unknown bias. The approach proposed by the ISO *Guide* to account for the uncertainty that arises from an unknown bias is now generally accepted. So we have used the ISO *Guide* to develop the following systematic laboratory-effects model.

2.4.1 Systematic Laboratory-Effects Model

We start with a combined result of the form $\sum_i a_i x_i$, where $\Sigma_i a_i = 1$, that is used as an initial estimate for Y. This estimate requires the assumption that Y is within the range of results $x_1, \ldots, x_n$. We refer to the initial estimate as the uncorrected combined result (UCR) and denote it by $x_{UCR} = \Sigma_i a_i x_i$. If $a_i = w_i / \Sigma_i w_i$, then x_{UCR} is the weighted mean $x_W = \Sigma_i w_i x_i / \Sigma_i w_i$, where $w_i = 1/u^2(x_i)$ for $i = 1, 2, \ldots, n$. If $a_i = 1/n$ for $i = 1, 2, \ldots, n$, then x_{UCR} is the arithmetic mean $x_A = \Sigma_i x_i / n$. Let $X_{UCR} = \Sigma_i a_i X_i$ be the expected value of the sampling distribution of x_{UCR}. According to *Assumption II*, the result x_{UCR} is subject to the bias $(X_{UCR} - Y)$. The ISO *Guide* recommends that the result x_{UCR} should be corrected to counter its possible bias and the uncertainty associated with the correction should be included in the combined standard uncertainty associated with the corrected result. The bias $(X_{UCR} - Y)$ is an unknown constant but the correction for bias, denoted by C, is a variable with a state-of-knowledge probability distribution. If the expected value and standard deviation of a state-of-knowledge probability distribution for the correction variable C are denoted by c and $u(c)$, respectively, then the correction applied to the result x_{UCR} to counter its possible bias is c and the standard uncertainty associated with the correction is $u(c)$.

[7] We did not introduce a new symbol for the weighted mean determined from model (2) because model (1) is a special case of model (2).

In order to specify a state-of-knowledge probability distribution for the correction variable C, the laboratory expected values $X_1, \ldots, X_n$ and the value Y of the measurand are regarded as variables with state-of-knowledge distributions and the data $x_1, \ldots, x_n$ and $u(x_1), \ldots, u(x_n)$ are regarded as given constants. A state-of-knowledge distribution for X_i represents the state of knowledge about the value Y of the measurand in the laboratory labeled i for $i = 1, 2, \ldots, n$. The expected value $E(X_i)$ and standard deviation $S(X_i)$ of the variable X_i are assumed to be x_i and $u(x_i)$, respectively, for $i = 1, 2, \ldots, n$ [5], [4]. It follows that $X_{\mathrm{UCR}} = \Sigma_i\, a_i X_i$ is a variable with a state-of-knowledge probability distribution. The expected value of X_{UCR} is $E(X_{\mathrm{UCR}}) = \Sigma_i\, a_i\, E(X_i) = \Sigma_i\, a_i\, x_i = x_{\mathrm{UCR}}$. In the expression $(Y - X_{\mathrm{UCR}})$ for the negative of bias, treated as a variable, we replace X_{UCR} with its expected value x_{UCR}. Then a probability distribution for C represents belief about the possible values of $(Y - x_{\mathrm{UCR}})$, where x_{UCR} is a constant and Y is the variable. The belief about possible values of Y is based on all available information including results of measurement and scientific judgment. In reference [6], we proposed a triangular distribution for the correction variable C, with peak at 0 and default limits $[x_{(1)} - x_{\mathrm{UCR}}] = \min\{x_1 - x_{\mathrm{UCR}}, \ldots, x_n - x_{\mathrm{UCR}}\}$ and $[x_{(n)} - x_{\mathrm{UCR}}] = \max\{x_1 - x_{\mathrm{UCR}}, \ldots, x_n - x_{\mathrm{UCR}}\}$. A criticism of the proposed triangular distribution with default limits is that it is determined by the extreme results $x_{(1)} = \min\{x_1, \ldots, x_n\}$ and $x_{(n)} = \max\{x_1, \ldots, x_n\}$, which are sometimes suspected to be in error.

Here, we propose a discrete-equal-probability distribution that is determined by all of the results $x_1, \ldots, x_n$. The results $x_1, \ldots, x_n$ are plausible values of Y as determined by competent laboratories.[8] So the known constant differences $(x_1 - x_{\mathrm{UCR}}), \ldots, (x_n - x_{\mathrm{UCR}})$ are plausible values of $(Y - x_{\mathrm{UCR}})$. These differences are a statistical basis for specifying a probability distribution for C. Let $c_i = x_i - x_{\mathrm{UCR}}$ for $i = 1, 2, \ldots, n$. Suppose $c_1, \ldots, c_n$ are assigned probabilities $p_1, \ldots, p_n$. Then the expected value of C is $c = E(C) = \Sigma_i\, p_i\, c_i = (\Sigma_i\, p_i\, x_i) - x_{\mathrm{UCR}}$ and the standard deviation of C is $u(c) = S(C) = \sqrt{[\Sigma_i p_i\, (c_i - c)^2]}$. Frequently, the available scientific knowledge is inadequate to assign different probabilities $p_1, \ldots, p_n$ to $c_1, \ldots, c_n$. Therefore, we propose the discrete-equal-probability distribution for which $p_i = 1/n$ for $i = 1, 2, \ldots, n$. The expected value and standard deviation of C based on discrete-equal-probability distribution are $c = x_{\mathrm{A}} - x_{\mathrm{UCR}}$ and $u(c) = \sqrt{[\Sigma_i\, (x_i - x_{\mathrm{A}})^2 / n]}$ respectively, where $x_{\mathrm{A}} = \Sigma_i\, x_i / n$ is the arithmetic mean of the results $x_1, \ldots, x_n$.

A measurement equation is required to incorporate correction for possible bias in a combined result of measurement for Y. The measurement equation that corresponds to the bias $(X_{\mathrm{UCR}} - Y)$ in the uncorrected combined result x_{UCR} is $Y = X_{\mathrm{UCR}} + C$. This measurement equation is widely applicable in metrology [10]. It suggests the following model for the value Y of the measurand:

$$E(X_i) = x_i,\ S(X_i) = u(x_i),\ X_{\mathrm{UCR}} = \sum_i a_i X_i,\ Y = X_{\mathrm{UCR}} + C, \tag{3}$$

where $a_1, \ldots, a_n$ are constants such that $\Sigma_i\, a_i = 1$. In this model, $X_1, \ldots, X_n$, X_{UCR}, C, and Y are variables with state-of-knowledge distributions. The expected value and standard deviation of X_i are the given constants x_i and $u(x_i)$, respectively, for $i = 1, 2, \ldots, n$. A state-of-knowledge distribution for the correction variable C is defined independently of the state-of-knowledge distributions for the variables $X_1, \ldots, X_n$, after the latter have been specified. In particular, X_{UCR} and C are independently distributed. We refer to model (3) [represented by Eq. (3)] as a systematic laboratory-effects model to distinguish it from the random laboratory-effects model (2) that regards the biases (systematic errors) $b_1, \ldots, b_n$ as random variables having the same sampling distribution with expected value zero. Suppose the standard deviation of the variable X_{UCR} is $S(X_{\mathrm{UCR}}) = u(x_{\mathrm{UCR}})$. Then the corrected combined result for Y determined from the systematic laboratory-effects model (3) is $y = x_{\mathrm{UCR}} + c$ and its associated standard uncertainty is $u(y) = \sqrt{[u^2(x_{\mathrm{UCR}}) + u^2(c)]}$.

The systematic laboratory-effects model (3) allows for the possibility that not all pairs of the variables $X_1, \ldots, X_n$ may be independently distributed. The variance $V(X_{\mathrm{UCR}}) = u^2(x_{\mathrm{UCR}})$ is determined from the variances and covariances of the variables $X_1, \ldots, X_n$. When the distributions of $X_1, \ldots, X_n$ are independent and X_{UCR} is the weighed mean $X_{\mathrm{W}} = \Sigma_i\, w_i X_i / \Sigma_i\, w_i$, where $w_i = 1/V(X_i) = 1/u^2(x_i)$ for $i = 1, 2, \ldots, n$, then $u^2(x_{\mathrm{UCR}}) = V(X_{\mathrm{W}}) = 1/[\Sigma_i\, w_i] = 1/\{\Sigma_i\, [1/u^2(x_i)]\}$. When the distributions of $X_1, \ldots, X_n$ are independent and X_{UCR} is the arithmetic mean $X_{\mathrm{A}} = \Sigma_i\, X_i / n$, then $u^2(x_{\mathrm{UCR}}) = V(X_{\mathrm{A}}) = (1/n^2)\Sigma_i\, V(X_i) = (1/n^2)\Sigma_i\, u^2(x_i)$[9].

In order to specify c and $u(c)$, one is free to use any reasonable distribution for C, based on scientific judgment. When the discrete-equal-probability distribution

[8] As noted in the footnote of Sec. 2.1, the total interval consisting of the union of intervals $[x_i \pm 2u(x_i)]$, for $i = 1, 2, \ldots, n$, represents an approximate range of the plausible values of Y.

[9] Since the harmonic mean of positive numbers is less than or equal to their arithmetic mean, $V(X_{\mathrm{W}}) \le V(X_{\mathrm{A}})$. When $u^2(x_1), \ldots, u^2(x_n)$ are equal, $V(X_{\mathrm{W}}) = V(X_{\mathrm{A}})$.

is used, $c = x_A - x_{UCR}$ and $u(c) = \sqrt{[\Sigma_i (x_i - x_A)^2/n]}$. In that case, the result of measurement for Y is $y = x_{UCR} + c = x_{UCR} + x_A - x_{UCR} = x_A$ and $u(y) = \sqrt{[u^2(x_{UCR}) + u^2(c)]}$, where $u(c) = \sqrt{[\Sigma_i (x_i - x_A)^2/n]}$.

Following the ISO *Guide*, the result y and uncertainty $u(y)$ determined from the systematic laboratory-effects model (3) are interpreted as the expected value and standard deviation of a state-of-knowledge distribution for the values that could reasonably be attributed to Y based on the data $x_1, \ldots, x_n$ and $u(x_1), \ldots, u(x_n)$ [5], [4], [6]. Thus the key comparison reference value x_R based on the systematic laboratory-effects model (3) is y and uncertainty $u(x_R)$ is $u(y)$. The corresponding degrees of equivalence are $d_i = x_i - y$ and $d_{i,j} = x_i - x_j$ for $i, j = 1, 2, \ldots, n$ and $i \neq j$. The uncertainties $u(d_i)$ and $u(d_{i,j})$ are determined from state-of-knowledge distributions for the variables $X_1, \ldots, X_n$ and Y.

3. Interpretation of the Key Comparison Reference Value and Its Associated Uncertainty

3.1 Classical Statistics Models Based on *Assumption I*

The nonexistent laboratory-effects model and the random laboratory-effects model are based on classical (frequentist) statistics. In particular, the results $x_1, \ldots, x_n$ are regarded as realizations of random variables with sampling distributions and Y is an unknown constant. Therefore, the key comparison reference value x_R is a realization of a random variable with a sampling distribution that has expected value Y and standard deviation $u(x_R) = u(x_W) = 1/\sqrt{[\Sigma_i w_i]}$. In the nonexistent laboratory-effects model w_i is $1/u^2(x_i)$ and in the random laboratory-effects model w_i is $1/[s_b^2 + u^2(x_i)]$ for $i = 1, 2, \ldots, n$. The interval $[x_R \pm 2u(x_R)]$ determined from a classical statistics model is a confidence interval for Y computed from the data $x_1, \ldots, x_n$ and $u(x_1), \ldots, u(x_n)$. Imagine that the CIPM key comparison could be repeated infinitely many times in exactly the same conditions using exactly the same instruments and artifacts. Now imagine that throughout these repetitions exactly the same sampling distributions continued to apply to the random variables $x_1, \ldots, x_n$. Then the confidence level is the fraction of the infinitely many hypothetical intervals, such as $[x_R \pm 2u(x_R)]$, that would include Y [4].

3.2 Systematic Laboratory-Effects Model Based on *Assumption II*

The key comparison reference value x_R and uncertainty $u(x_R)$ determined from the systematic laboratory-effects model are given constants that represent the expected value and standard deviation of a state-of-knowledge distribution for Y based on the data $x_1, \ldots, x_n$ and $u(x_1), \ldots, u(x_n)$. The interval $[x_R \pm 2u(x_R)]$ determined from the systematic laboratory-effects model is an expanded uncertainty interval for Y. The coverage probability (level of confidence) of the interval $[x_R \pm 2u(x_R)]$ is the fraction of a state-of-knowledge distribution for Y that is encompassed by this interval [4].

4. Interpretation of the Degrees of Equivalence and Their Associated Uncertainties

4.1 Classical Statistics Models Based on *Assumption I*

In the random laboratory-effects model and its special case the nonexistent laboratory-effects model, the expected values of the sampling distributions of $x_1, \ldots, x_n$, and x_R are all equal to Y. Therefore, the expected values of the sampling distributions of all degrees of equivalence $d_i = x_i - x_R$ and $d_{i,j} = x_i - x_j$ are zero, for $i, j = 1, 2, \ldots, n$ and $i \neq j$. This implies that all computed degrees of equivalence, whether small or large, are statistical estimates of zero. In particular, according to these models, all degrees of equivalence published in the key comparison database (KCDB) [11] are estimates of zero.

4.2 Systematic Laboratory-Effects Model Based on *Assumption II*

In the systematic laboratory-effects model, the results $x_1, \ldots, x_n$ are the expected values and the uncertainties $u(x_1), \ldots, u(x_n)$ are the standard deviations of state-of-knowledge distributions for the laboratory expected values $X_1, \ldots, X_n$, treated as variables. It follows that the degree of equivalence $d_i = x_i - x_R = x_i - y$ is the expected value of a state-of-knowledge distribution for the laboratory effect (bias) $X_i - Y$ for $i = 1, 2, \ldots, n$, and the degree of equivalence $d_{i,j} = x_i - x_j$ is the expected value of a state-of-knowledge distribution for the difference $X_i - X_j$ for $i, j = 1, 2, \ldots, n$ and $i \neq j$. The uncertainty $u(d_i)$ is the standard deviation[10] of $X_i - Y$ and the uncertainty $u(d_{i,j})$ is the standard deviation of $X_i - X_j$, for $i, j = 1, 2, \ldots, n$ and $i \neq j$.

[10] The standard deviation of $X_i - Y$ depends on the covariance between X_i and Y for $i = 1, 2, \ldots, n$. Since $Y = X_{UCR} + C = \Sigma_i a_i X_i + C$ and the variable C is distributed independently of the variables $X_1, \ldots, X_n$, the covariances $C(X_i, Y)$, for $i = 1, 2, \ldots, n$, can be determined from the variances and covariances of $X_1, \ldots, X_n$. Then $u(d_i) = \sqrt{[V(X_i - Y)]}$, where the variance $V(X_i - Y)$ is equal to $V(X_i) + V(Y) - 2 \times C(X_i, Y)$.

5. Conclusion

We addressed a simple CIPM key comparison where the common measurand is a physical quantity of stable value during the comparison. We discussed statistical interpretation of the key comparison reference value, the degrees of equivalence, and their associated uncertainties determined from the following three statistical models: nonexistent laboratory-effects model, random laboratory-effects model, and systematic laboratory-effects model. The first two models are based on classical (frequentist) interpretation of measurements. The systematic laboratory-effects model is based on Bayesian interpretation of measurements.

The key comparison reference value x_R and uncertainty $u(x_R)$ determined from the systematic laboratory-effects model represent the expected value and standard deviation of a state-of-knowledge distribution for the value Y of the measurand. Therefore their statistical interpretation agrees with the ISO *Guide*. According to the systematic laboratory-effects model, the degree of equivalence d_i and uncertainty $u(d_i)$ are, respectively, the expected value and standard deviation of a state-of-knowledge distribution for the laboratory effect (bias) $X_i - Y$, for $i = 1, 2, \ldots, n$, and the degree of equivalence $d_{i,j}$ and uncertainty $u(d_{i,j})$ are, respectively, the expected value and standard deviation of a state-of-knowledge distribution for the difference $X_i - X_j$, for $i, j = 1, 2, \ldots, n$ and $i \neq j$. Thus the degrees of equivalence determined from the systematic laboratory-effects model quantitate the agreements and disagreements of laboratory results. Therefore, the systematic laboratory-effects model is suitable for the data analysis of a simple CIPM key comparison.

Acknowledgment

The following provided helpful comments on earlier drafts of this paper: T. V. Vorburger, Ron Boisvert, Eric Shirley, Tony Kearsley, Jim Gardener, and Nell Sedransk.

6. References

[1] Mutual recognition of national measurement standards and of calibration and measurement certificates issued by national metrology institutes, International Committee of Weights and Measures (CIPM), 14 October 1999, (http://www1.bipm.org/utils/en/pdf/mra_2003.pdf).

[2] Guidelines for CIPM key comparisons, International Committee of Weights and Measures (CIPM), 1 March 1999, (http://www1.bipm.org/utils/en/pdf/guidelines.pdf).

[3] L. Nielsen, Evaluation of measurement intercomparisons by the method of least squares, Report DFM-99-R39, 3208 LN, Danish Institute of Fundamental Metrology, Lyngby, Denmark (2000).

[4] R. N. Kacker and A. T. Jones, On use of Bayesian statistics to make the Guide to the Expression of Uncertainty in Measurement consistent, Metrologia **40**, 235-248 (2003).

[5] Guide to the Expression of Uncertainty in Measurement, 2nd Ed., Geneva, International Organization for Standardization, ISBN 92-67-10188-9 (1995).

[6] R. N. Kacker, R. U. Datla, and A. C. Parr, Combined result and associated uncertainty from interlaboratory evaluations based on the ISO Guide, Metrologia **39**, 279-293 (2002).

[7] S. R. Searle, Linear Models, 1971, John Wiley & Sons Inc., New York.

[8] R. N. Kacker, Combining information from interlaboratory evaluations using a random effects model, Metrologia **41**, 132-136 (2004).

[9] R. C. Paule and J. Mandel, Consensus values and weighting factors, J. Res. Natl. Bur. Stand.(U.S.) **87**, 377-385 (1982).

[10] I. Lira and W. Wöger, Bayesian evaluation of the standard uncertainty and coverage probability in a simple measurement model, Meas. Sci. Technol. **12**, 1172-1179 (2001).

[11] BIPM key comparison data base, (http://kcdb.bipm.org).

About the authors: *Dr. R. N. Kacker is a mathematical statistician in the Mathematical and Computational Sciences Division of the NIST Information Technology Laboratory. Dr. R. U. Datla and Dr. A. C. Parr are physicists in the Optical Technology Division of the NIST Physics Laboratory. Dr. Parr is the division chief and Dr. Datla is a group leader. The National Institute of Standards and Technology is an agency of the Technology Administration, U.S. Department of Commerce.*